

AD-A111 014

CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS F/O 12/1  
THE CONTRIBUTIONS OF WILLIAM COCHRAN TO CATEGORICAL DATA ANALYSIS-ETC(U)  
DEC 81 S E FIENBERG N00014-80-C-0637  
UNCLASSIFIED TR-222 NL

101  
21008

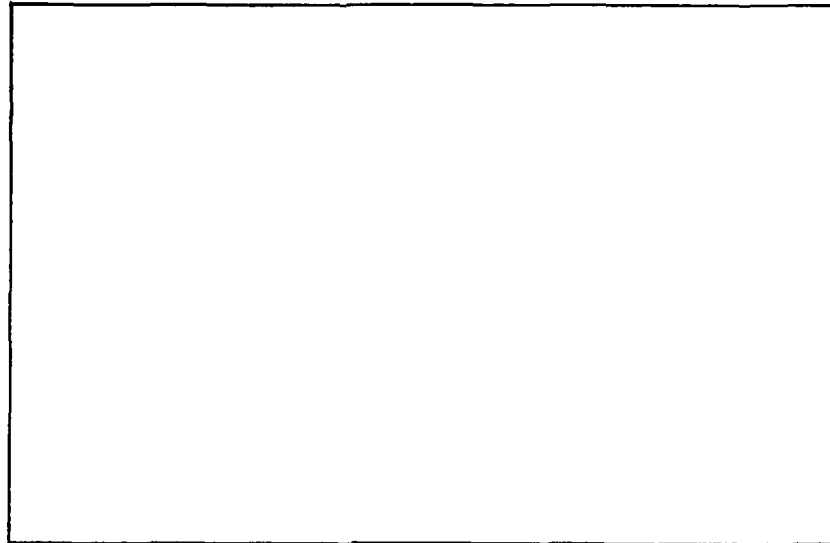
END  
DATE  
FILMED  
19-82  
DTIC

AD A111014

LEVEL

II

①



12 24

DEPARTMENT  
OF  
STATISTICS

DEC  
FEB 17 1932  
E

DTC FILE COPY

391190

**Carnegie-Mellon University**

PITTSBURGH, PENNSYLVANIA 15213

This document is loaned to you for public release and its distribution is unlimited.

82 02 17 062

75

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
/or	
Dist _____	
<b>A</b>	



# THE CONTRIBUTIONS OF WILLIAM COCHRAN TO CATEGORICAL DATA ANALYSIS

by

Stephen E. Fienberg

Department of Statistics  
Carnegie-Mellon University  
Technical Report No. 222  
December, 1981

The preparation of this paper was partially supported by the Office of Naval Research Contract N00014-80-C-0637 at Carnegie-Mellon University. Reproduction in whole or part is permitted for any purpose of the United States Government. Thanks are due to S. May for computational assistance, and to J. Tanur for comments on an earlier draft.

This document is approved  
for public release and  
distribution in its present form.

## 1. Introduction

The key words in the title of this paper are "data analysis," for in the categorical data area, as in almost all of his research work, William G. Cochran was concerned with the practical aspects of statistical methods and theory. Although his early papers on the topic (e.g. Cochran 1936a, 1940) included formal mathematical statistical derivations, their focus was always on actual applications and on ways of adjusting theoretical results to deal with the data at hand. This early practical orientation extends throughout Cochran's research on the statistical analysis of categorical data, research that spanned five decades.

This review of Cochran's contributions to categorical data analysis focuses on what I take to be his four major contributions:

- (a) the distribution of  $\chi^2$  test statistics in the presence of small expectations, and the correction for continuity,
- (b) the Q-test for comparing percentages in matched samples,
- (c) methods for strengthening  $\chi^2$ -tests,
- (d) the Cochran test for combining results from several  $2 \times 2$  tables.

Each of these contributions, by itself, would have been sufficient to establish any ordinary statistician's claim as a statistical innovator. Taken together they represent pathbreaking work by an extraordinary statistician, who was, at the same time, making major contributions to other areas of statistics, such as the design of experiments and sample surveys.

Although the contributions listed above were all published in the 1940's and 1950's, Cochran continued to monitor the progress of research on categorical data analysis, especially during the renaissance of the mid-1960's, out of which came new methods for the analysis of loglinear and logit models. I suspect this continuing interest stemmed from his understanding the practical importance of this work for other problems in which he was interested. I remember Bill approaching me in 1967 to discuss his reply to a letter from Berkson that dealt with simultaneous logit equations for a trichotomous response. In later years, we had several

conversations about the work I was doing with Fred Mosteller, Yvonne Bishop, and Paul Holland on loglinear models, and in Chapter 20 of the 7th edition of Snedecor and Cochran (1980). Bill added several paragraphs on the loglinear model, and how it provides a generalization for the logit methods described there (and in the earlier 6th edition).

## 2. The Distribution of $\chi^2$ Test Statistics in Presence of Small Expected Values

In his first paper on the analysis of categorical data (1936a), written while he was a member of the staff of the Rothamsted Experiment Station, Cochran presents a detailed analysis of the results of an experiment focussing on the distribution of diseased plants in a 4x4 Latin Square layout of plots, each with 6 rows containing 15 tomato plants. After beginning with an actual field map of the diseased plants, he presents the variance test for the homogeneity of  $N$  binomial proportions, showing (using an argument suggested by Fisher) that it is identical to the usual  $\chi^2$  test for homogeneity in a  $2 \times N$  table. In applying this test to the tomato plant data, Cochran expresses concern about the effect of small expectations on the  $\chi^2$  approximation for the distribution of the test statistic, and he illustrates how one can study the effect by looking at the exact distribution of  $\chi^2$  (something that is not explained in detail until the second 1936 paper). He then illustrates the results of  $N = 10$  groups of  $n = 9$  plants each where the overall disease rate is  $p = 0.2$ . The expectations are 1.8 and 7.2 (the former was considered quite small back in 1936), yet the highest percentage discrepancy between the exact probability and the  $\chi^2$  approximation is about 16% at around  $p = .2$ . The  $\chi^2$  approximation is surprisingly good, especially at  $p = .05$  and  $p = .01$ .

The remainder of this first 1936 paper focusses on supplementary analyses, such as an analysis of variance on the number of diseased plants to take into account the row and column structure (this is a section of the paper that might have benefited from the modern methods for logit models), and the issue of contagion of disease using a test for runs. All in all, Cochran presents an illuminating example of data analysis, and, in the process of doing so, he introduces several technical proofs of results, as well as a first attempt at dealing with the

problem of small expectations.

In what appears to be a companion piece to the first paper, Cochran (1936b) focusses directly on the effect of small expectations on the binomial variance test (and thus the  $\chi^2$  test for  $2 \times N$  tables). Here we learn that by "the exact distribution of  $\chi^2$ ," Cochran means the distribution of  $\chi^2$  based on the conditional distribution given both the row and column totals in the  $2 \times N$  table. That is, suppose  $x_1, x_2, \dots, x_N$  are the number of successes out of  $n$  for  $N$  binomials, each with probability of success  $p$ . Then Cochran looks at the conditional distribution of  $x_1, x_2, \dots, x_N$  given the total  $T = x_1 + x_2 + \dots + x_N$ , i.e. at the probability

$$P' = \frac{(n!)^N T! (Nn - T)!}{(Nn)! \prod_i [x_i! (n - x_i)!]} \quad (2.1)$$

Because the  $x_i$ 's are typically not all distinct the probability in (2.1) corresponds to  $N! / \prod_i a_i!$  possible arrangements of the  $N$   $x_i$ 's, where there are  $a_0$  0's,  $a_1$  1's, etc. Thus the exact probability is

$$P = \frac{N!}{\prod_i a_i!} P' \quad (2.2)$$

The exact distribution of  $\chi^2$  then comes from an enumeration of all possible sample configurations, and the calculation of the probability in (2.2) and the corresponding value of

$$\chi^2 = \sum_{i=1}^N \frac{(x_i - T/N)^2}{T(nN - T)/N^2 n} \quad (2.3)$$

Cochran never presents any discussion of or formal justification for why the relevant probability of interest should be based on the conditional distribution given the sufficient statistic  $T$ , although presumably he was strongly influenced by Fisher and Yates on this issue. (For a modern-day justification of the use of the conditional distribution, with which I am not in total sympathy, see Cox, 1970, pp. 44-46.).

Cochran explains this second 1936 paper by taking us through a detailed example for  $n = 8$ ,  $p = .25$ , and  $N = 4$ . Although he does not explain it in the text, he appears to have taken  $T = Nnp = 8$  (somewhat arbitrarily). A summary table includes 13 of the 15 possible values of the  $\chi^2$  statistic (the two highest values being omitted), the corresponding exact tail probabilities ( $P$  as in (2.2)), the tabular values from the  $\chi^2$  distribution, and an adjusted tabular  $\chi^2$  value corrected for continuity (more on this later). In this and four other examples, Cochran finds tolerable percentage discrepancies in the range  $0.1 \geq P \geq 0.005$  resulting from the use of the tabular  $\chi^2$  values, even when  $np$  is as low as 0.9, and a tendency for the adjusted tabular  $\chi^2$  values to overestimate the exact tail values,  $P$ .

For the special case of  $np$  and  $n$  small but  $N$  large, Cochran investigates the normal approximation, using expansions for the mean and variance, and he finds a result somewhat different from the usual normal approximation to the  $\chi^2$  distribution. Here and earlier he investigates not only the binomial problem but also the Poisson limit in which  $n \rightarrow \infty$ ,  $p \rightarrow 0$ , and  $np \rightarrow m$ .

The final section of the second 1936 paper recalls Fisher's justification of the validity of the  $\chi^2$  statistic, i.e. that it approximates minus twice the loglikelihood ratio. For the Poisson limit, this latter quantity is

$$G^2 = 2 \sum_i x_i \log \frac{x_i}{m_i} \quad (2.4)$$

We note that expression (2.4) can be of use as a generic form for the loglikelihood ratio statistic that parallels the generic Pearson  $\chi^2$  formula  $X^2 = \sum_i (x_i - m_i)^2 / m_i$ . Cochran then returns to his example with  $m = 2$  and  $N = 4$  and shows that the tabular  $\chi^2$  values, both corrected and uncorrected, *underestimate* the tail probabilities calculated from the exact distribution for  $G^2$ , although the agreement is "not sensibly worse" than that for the Pearson statistic,  $X^2$ .

In this pair of 1936 papers we see Cochran's attempt to grapple with the difficult practical problem presented by small expectations using a mixture of mathematical statistics theory and tabular calculations. These papers leave many practical and theoretical questions unanswered, and Cochran returned to some of these in his subsequent research.

In a 1942 paper, for example, he focussed his attention on the correction for continuity for  $\chi^2$  tests, giving a careful statement of how the  $\chi^2$  distribution approximation comes about if one regards "the exact discontinuous distribution as a grouping of the tabular distribution, with each possible value of the exact  $\chi^2$  representing all values of the continuous  $\chi^2$  which are nearer to it than to any other permissible value of  $\chi^2$ ." Using this formulation, Cochran is able to state a general rule in applying corrections for continuity:

Calculate  $\chi^2$  by the usual formula. Find the next lowest possible value of  $\chi^2$  to the one to be tested, and use the tabular probability for a value of  $\chi^2$  midway between the two.

If the possible values of  $\chi^2$  are closely spaced together, the probabilities given by the uncorrected  $\chi^2$  and the corrected  $\chi^2$  may differ only by an amount that is regarded as negligible. In this case the correction may be ignored.

He notes that this general rule, when applied to  $\chi^2$  for the 2x2 table, results in the usual Yates correction.

Cochran then turns to the special problem which occurs when a  $\chi^2$  statistic can be partitioned into single-degree-of-freedom components. He argues both on theoretical grounds and numerically that, if the total  $\chi^2$  is to be computed as the sum of its components, it is appropriate to add uncorrected  $\chi^2$  values and then correct the total using the general rule, if necessary. Adding corrected values badly overcorrects, and the agreement with the exact distribution gets steadily worse as the number of  $\chi^2$  components grows.

At the end of this 1942 paper, Cochran returns to the issue of small expectations, and develops a new approach to exploring the effect of one small expected value. He begins with a trinomial problem and takes three examples with  $n = 20$ , where  $m_1/m_2 = 3/2$  but  $m_3$  varies from 1.0 to 0.1. If we treat the  $m_i$ 's as known, then



$$\chi^2 = \sum_{i=1}^3 \frac{(x_i - m_i)^2}{m_i} \quad (2.5)$$

has an asymptotic  $\chi^2$  distribution with 2 degrees of freedom. Cochran compares the exact distribution with the tabular  $\chi^2$  values, once again finding fairly good agreement down to about  $p = .03$  for both  $m_3 = 1$  and  $m_3 = .5$ . But for  $m_3 = .1$  he notes that "the tabular  $\chi^2$  is useless as an approximation throughout the whole of the region between  $p = .1$  and  $p = .01$ ." Because of the serious underestimation of the tail area probabilities for  $m_3 = .1$ , and beyond  $p = .03$  for  $m_3 = .5$  and  $m_3 = 1$ , Cochran suggests that a correction for continuity cannot repair the discrepancy, and thus discontinuity is not its principal cause. This leads him to approximate the exact tail area for the  $r$ -cell multinomial problem in the presence of a small expectation in the  $r$ th cell by

$$\chi^2_{r-2} + \frac{(k - m_r)^2}{m_r} \quad (2.6)$$

where  $k$  is Poisson with mean  $m_r$ . This approximation fits extremely well for the trinomial exact distribution with  $m_3 = .1$ . Finally Cochran uses this approximation to come up with the following recommendations for the minimum degrees of freedom,  $r - 1$ , to ensure that the usual  $\chi^2$  approximation is accurate to within 20% at the .05 and .01 level, for various values of  $m_i$ :

Smallest expectation	0.1	0.5	1.0	2.0	3.0	5.0
Minimum d.f.	?	25	10	6	4	2

The paper concludes with a brief look at the problem of two small expectations.

A decade later, Cochran (1952, 1954) was still refining his advice on the minimum expectations to be used in  $\chi^2$  tests. The following recommendations from Cochran (1954) are reasonably consistent with his earlier research described above, but were based on some additional but unpublished calculations:

- (a) *Goodness of fit tests of unimodal distributions (such as the normal or Poisson).*  
Here the expectations will be small only at one or both tails. Group so that the

minimum expectation at each tail is at least 1.

- (b) *The 2 X 2 table* Use Fisher's exact test (i) if the total  $N$  of the table  $< 20$ , (ii) if  $20 < N < 40$  and the smallest expectation is less than 5. . . . If  $N > 40$  use  $\chi^2$ , corrected for continuity.
- (c) *Contingency tables with more than 1 d.f.* If relatively few expectations are less than 5 (say in 1 cell out of 5 or more, or 2 cells out of 10 or more), a minimum expectation of 1 is allowable in computing  $\chi^2$ .

Contingency tables with most or all expectations below 5 are harder to prescribe for. With very small expectations, the exact distribution of  $\chi^2$  can be calculated without too much labor. . . . If  $\chi^2$  has less than 30 degrees of freedom and the minimum expectation is 2 or more, use of the ordinary  $\chi^2$  tables is usually adequate. If  $\chi^2$  has more than 30 degrees of freedom, it tends to become normally distributed, but when the expectations are low, the mean and variance are different from those of the tabular  $\chi^2$ . . . . Compute the exact mean and variance, and treat  $\chi^2$  as normally distributed with that mean and variance.

A hidden issue in all of Cochran's work on  $\chi^2$  tests in the presence of small expectations is the relevant reference distribution for the test statistic. Cochran consistently uses the "exact distribution" of  $X^2$ , conditional on the values of the margins, as his reference distribution. This was consistent with the work of Fisher and Yates, and allowed Cochran actually to compute distributions in the pre-computer era because he needed to worry about only a relatively small number of possible values for the test statistic. An alternative to Cochran's approach is to use the large-sample  $\chi^2$  distribution rather than the exact distribution as the reference distribution. If one of our aims is to correct the  $\chi^2$ -statistic to conform more closely to the actual  $\chi^2$  distribution, the usual correction for continuity, suggested for use by Cochran, results in an overly conservative test (see e.g. Plackett, 1964; Grizzle, 1967; and Conover, 1974). Moreover, the discrepancy in behavior between the  $X^2$  and  $G^2$  statistics from this large-sample perspective can be attributed to the differing influence of cells with very small observed as opposed to expected counts (see Larntz, 1978; Fienberg, 1979).

### 3. The Q-Test for Matched Proportions

Cochran's (1950) paper on the comparison of percentages in matched samples is a gem. He begins with the case of two matched proportions and presents the well-known McNemar-Mosteller test. He then generalizes this test statistic to the case where there are  $c > 2$

matched samples. The setup is as follows.

Consider an  $n \times c$  table listing the results on  $c$  binary items (the  $c$  matched samples) for  $n$  individuals. Let  $x_{ij}$  be 1 if there is a success for individual  $i$  on item  $j$ , and assume that the total number of successes in each row is fixed (i.e.:

$$u_i = \sum_{j=1}^c x_{ij} \quad (3.1)$$

is fixed for  $i = 1, 2, \dots, n$ ). Cochran proposes a test statistic based on randomization of zeros and ones within rows. Since rows containing only 0's or only 1's (i.e. rows for which  $u_i = 0$  or  $u_i = c$ ) aren't changed under such a randomization, they play no part in Cochran's test statistic, and thus they can be eliminated from consideration. The array  $\{x_{ij}\}$  is thus reduced to size  $r \times c$ , where the row total  $u_i$  can take values  $1, 2, \dots, c-1$ .

Cochran's  $Q$ -test *for the equality of the proportions* in the matched samples is then based on the sum of squared deviations of the  $c$  column totals,  $T_j$ , of the  $r \times c$  table, suitably normalized so that the resulting quantity follows an asymptotic  $\chi^2$  distribution:

$$Q = \frac{c(c-1)\sum(T_j - T)^2}{c(\sum u_i) - (\sum u_i^2)} \quad (3.2)$$

Under the randomization distribution with fixed row totals, the asymptotic distribution of  $Q$  is  $\chi^2$  with  $c-1$  d.f., and Cochran presents a heuristic binomial-like argument leading to this result. For the case  $c=2$ ,  $Q$  reduces to the McNemar test statistic. Cochran's key step in the  $c > 2$  case is the use of the randomization distribution, but this procedure actually specifies a null hypothesis that is far more restrictive than the null hypothesis of interest, i.e. equality of proportions for the  $c$  items.

An alternative way to view the matched proportions problem is to treat each row of the  $\{x_{ij}\}$  array as an independent observation from a  $2^c$  contingency table. The hypothesis of equality of matched proportions can then be translated into the hypothesis of homogeneity of

the one-way marginal proportions in the  $2^k$  table. The randomization distribution used to derive the Q statistic for the  $\{x_{ij}\}$  array implies complete symmetry (under permutation of dimensions) for the  $2^k$  table. As Bishop, Fienberg, and Holland (1975, Chapter 8) note, Cochran's Q-test is then equivalent to a likelihood ratio test for the model of complete symmetry (or one-way marginal homogeneity) given the model of one-way quasi-symmetry. Although Cochran was quite explicit about the complete symmetry implicit in the randomization justification of the Q test, many of those who subsequently investigated the properties of the test failed to understand its implications.

There is another interesting aspect of the Q-test which links it to the modern literature on loglinear models for contingency table analysis. Plackett (1981), in a very brief section on the 2nd edition of his monograph on the analysis of categorical data, notes that Cochran's Q-statistic can also be viewed as a means for testing the equality of item parameters in the Rasch (1980) model. That is, if

$$\log \frac{P(x_{ij}=1)}{P(x_{ij}=0)} = \psi_j - \mu - v_i \quad (3.3)$$

then Q can be used to test

$$H: v_j = 0 \quad \text{for all } j. \quad (3.4)$$

This observation, although it seems simple on the surface, is directly related to new and exciting ideas for the analysis of the Rasch model using loglinear models applied to the corresponding  $2^k$  contingency table, that have recently been proposed by Duncan (1982), and by Tjur (1981).

In the 1950 paper, Cochran also investigates, for 8 specific cases, the small sample behavior of Q by comparing the  $\chi^2$  approximations (with and without a correction for continuity based on the general rule he had proposed in 1942 with the exact distribution based on randomization. He also introduces an F approximation which fares poorly. As a result of

these comparisons. Cochran concludes that the use of the uncorrected  $\chi^2$  approximation is superior, and quite adequate for most situations to which the Q statistic should be applied.

In retrospect, what makes the 1950 paper on matched proportions so remarkable is how complete and how innovative it was, on the one hand, and yet how many problems it opened up for other statisticians to investigate, subsequently, on the other. This combination was a characteristic of Cochran's writings: he was always thorough, but he also pointed out aspects of problems worthy of further investigation, and encouraged others to pursue them.

#### 4. Methods for Strengthening $\chi^2$

In another remarkable paper, Cochran (1954) provided an incisive review of methods for strengthening or supplementing the common uses of the  $\chi^2$  test, and at the same time he presented two new and important classes of tests. The first of these, discussed in this section, is a class of 1 d.f. tests for linear functions of the deviations between observed and expected counts. (A full justification of these tests, as well as a generalization to multiple degree-of-freedom tests, was subsequently given in Cochran (1955).) The second new class of tests was for combining results from several  $2 \times 2$  tables. Because of the importance of the latter, we treat it as a separate topic in Section 5.

Cochran begins the 1954 paper by noting that:

In this paper I want to discuss two kinds of failure to make the best use of  $\chi^2$  tests which I have observed from time to time in reading reports of biological research. The first arises because  $\chi^2$  tests, as has often been pointed out, are not directed against any specific alternative to the null hypothesis. . . . No attempt is made to detect any particular pattern of deviations ( $f_{ij} - m_{ij}$ ) that may hold if the null hypothesis is false. One consequence is that the usual  $\chi^2$  tests are often insensitive, and do not indicate significant results when the null hypothesis is actually false. Some forethought about the kind of alternative hypothesis that is likely to hold may lead to alternative tests that are more powerful and appropriate. Further, when the ordinary  $\chi^2$  test does give a significant result, it does not direct attention to the way in which the null hypothesis disagrees with the data, although the pattern of deviations may be informative and suggestive for future research. The remedy here is to supplement the ordinary test by additional tests that help to reveal the significant type of deviation.

The remedy for both kinds of failure is often the use of a single degree of freedom, or a group of degrees of freedom, from the total  $\chi^2$ , and Cochran explores such remedies for (a) the goodness-of-fit test for the Poisson distribution, the binomial distribution, and the normal distribution, and (b) two-way contingency tables.

For the distributional goodness-of-fit tests, Cochran presents new results for linear functions

$$L = \sum g_i (f_i - \hat{m}_i) \quad (4.1)$$

where  $f_i$  is the observed frequency,  $\hat{m}_i$  the estimated expected frequency in the  $i$ th cell, and the  $g_i$  are weights selected in advance in a way to make  $L$  sensitive to some likely alternative hypothesis. Then

$$\chi^2 = \frac{L^2}{\widehat{\text{Var}}(L)} \quad (4.2)$$

is asymptotically distributed as  $\chi^2$  with 1 d.f. when the distribution being fitted is appropriate. In applying this test in the specific cases of the Poisson, binomial, and normal, Cochran presents, in particular, the test statistic for a single deviation. He is careful to discuss the multiplicity problem that results from applying such tests to abnormally large deviations, after examining the data, and he suggests a Bonferroni approach to the calculation of significance levels in such instances.

Cochran then turns to the contingency table problem, focussing on the subdivision of degrees of freedom in the  $2 \times N$  case with fixed column totals, i.e.  $N$  binomials with probabilities of success,  $p_i$ ,  $i = 1, 2, \dots, N$ , and totals  $n_i$ ,  $i = 1, 2, \dots, N$ . He first looks at a division of the  $N$  columns into  $N_1$  and  $N_2 = N - N_1$ , and subdivides  $\chi^2$  into 3 components: (i) comparing the overall proportion of success in the first  $N_1$  columns with that in the last  $N_2$  (1 d.f.), (ii) comparing variation of proportions within the first  $N_1$  columns ( $N_1 - 1$  d.f.), (iii) comparing variation of proportions within the last  $N_2$  columns ( $N - N_1 - 1$  d.f.). The subdivision he suggests gives an additive partition of  $\chi^2$ , but Cochran notes that when the 1 d.f. component, (i), is

significant, this partition needs to be supplemented by direct  $\chi^2$ -tests for (ii) and (iii), thereby yielding a non-additive partition.

Other 1 d.f. components in the  $2 \times N$  table which Cochran examines include (a) a 1 d.f. test for linear regression of the proportions on some auxiliary variable,  $z$ ; (b) a 1 d.f. comparison of mean scores for the auxiliary variable (which turns out to be identical to (a)); and (c) a sequence of  $N-1$  cumulative 1 d.f. comparisons of the column proportions. For the latter tests, Cochran suggests partitioning the sum of squares in the numerator of  $\chi^2$  into

$$\frac{n_2 \{n_1 \hat{p}_1 - n_1 \hat{p}_2\}^2}{n_1 (n_1 + n_2)} \quad (4.3)$$

$$\frac{n_3 \{n_1 \hat{p}_1 + n_2 \hat{p}_2 - (n_1 + n_2) \hat{p}_3\}^2}{(n_1 + n_2)(n_1 + n_2 + n_3)} \quad (4.4)$$

and for the general term,

$$\frac{n_{r+1} \{n_1 \hat{p}_1 + \dots + n_r \hat{p}_r - (n_1 + \dots + n_r) \hat{p}_{r+1}\}^2}{(n_1 + \dots + n_r)(n_1 + \dots + n_{r+1})} \quad (4.5)$$

Note that (4.3) is based on the  $2 \times 2$  table:

$x_1$	$x_2$
$n_1 - x_1$	$n_2 - x_2$
$n_1$	$n_2$

(4.6)

and that each subsequent component cumulates the entries in the columns already examined and compares them with those in the next column. Thus the general term in (4.5) is based on

$\sum_{i=1}^r x_i$	$x_{r+1}$
$\sum_{i=1}^r (n_i - x_i)$	$n_{r+1} - x_{r+1}$
$\sum_{i=1}^r n_i$	$n_{r+1}$

(4.7)

Cochran's additive partition into  $(N-1)$  1 d.f.  $\chi^2$  components involves dividing each of (4.3), (4.4), (4.5), etc. by  $\hat{p}(1-\hat{p})$ , where

$$\hat{p} = \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N n_i} . \quad (4.8)$$

Cochran also mentions Lancaster's *non-additive* partition of  $\chi^2$ , which uses as the denominator for the  $r$ th component,  $\hat{p}^{(r-1)}(1-\hat{p}^{(r-1)})$ , where

$$\hat{p}^{(r-1)} = \frac{\sum_{i=1}^{r-1} x_i}{\sum_{i=1}^{r-1} n_i} . \quad (4.9)$$

This approach yields the usual 1 d.f.  $\chi^2$  tests for each of the  $2 \times 2$  tables of the form (4.7). It is interesting to observe that Cochran failed to note that, if likelihood ratio tests are substituted for  $\chi^2$  tests in each of the  $2 \times 2$  tables, the Lancaster partition is, in fact, additive (Kullback, 1959). This likelihood ratio partitioning of a  $2 \times N$  table leads rather naturally to the following alternative 1 d.f. test for an increasing (or decreasing) shift in the  $p_i$ 's, suggested recently by Stewart (1981).

The  $N-1$   $2 \times 2$  tables of the form (4.7) can be put together to form a  $2 \times 2 \times (N-1)$  three-way contingency table. The likelihood ratio test for independence in the original  $2 \times N$  table is identical to the test for conditional independence with the  $N-1$   $2 \times 2$  tables (this is the consequence of the Kullback partitioning result). Call the test statistic  $G_1^2$ . If we let the no-2nd-order interaction likelihood ratio test statistic be  $G_2^2$ , then

$$\Delta G^2 = G_1^2 - G_2^2 \quad (4.10)$$

is a 1 d.f. component of  $G_1^2$  which is sensitive to shift alternatives of the form:

$$p_2 = \frac{p_1}{(1-p_1)a + p_1} . \quad (4.11)$$



$$p_3 = \frac{n_1 p_1 + n_2 p_2}{(n_1 + n_2 - n_1 p_1 - n_2 p_2)c + n_1 p_1 + n_2 p_2} \quad (4.12)$$

and for the general term

$$p_{r+1} = \frac{\sum_{j=1}^r n_j p_j}{(\sum_{j=1}^r n_j - \sum_{j=1}^r n_j p_j)c + \sum_{j=1}^r n_j p_j} \quad (4.13)$$

where  $c$  is the constant odds-ratio implied by the no-second-order interaction model.

To illustrate this new 1 d.f. test we use Cochran's own example, the data for which we reproduce here as Table 4.1.

		Clinical Change				
		Marked	Moderate	Slight	Same	Worse
Degree of	0-7	11	27	42	53	11
Infiltration	8-15	7	15	16	13	1

Table 4.1

Data on Clinical Change by Degree  
of Infiltration [Cochran (1954, p.435)]

The standard tests for independence applied to the data of Table 4.1 yield

$$X^2 = 6.87$$

$$G^2 = 7.28$$

each with 4 d.f. The value of Cochran's 1 d.f.  $X^2$  test for regression is 6.67 (where the  $z$ 's for the columns are -3, -2, -1, 0, and 1).

Next we examine the data as a series of 4 2x2 tables, with columns cumulated from the left

as illustrated in Table 4.2. A test for conditional independence of degree of infiltration and clinical change, given "layer," yields

$$X^2_1 = 6.64$$

$$G^2_1 = 7.28$$

with 4 d.f. Note that  $G^2_1$  takes the identical value as that for the likelihood ratio test in the original table (numerical confirmation of Kullback's result). The test for no-2nd-order interaction yields

$$G^2_2 = 1.61$$

with 3 d.f., and thus the 1 d.f. component of  $G^2_1$  which tests for Stewart's shift alternative in (4.13) is

$$\Delta G^2 = 7.28 - 1.61 = 5.67.$$

This value is quite close to that for the regression 1 d.f.  $\lambda^2$  test, as we might expect.

Cochran would surely have found this new 1 d.f. alternative test quite appealing, especially given its indirect relationship to his own test for combining 2x2 tables, discussed in the following section.

Layer	Degree of Infiltration	Clinical Change	
		First Proportion	Second Proportion
1. (column 1 vs. column 2)	0-7	11	27
	8-15	7	15
2. (columns 1 and 2 vs. column 3)	0-7	38	42
	8-15	22	16
3. (columns 1,2. and 3 vs. column 4)	0-7	80	53
	8-15	38	13
4. (columns 1,2,3. and 4 vs. column 5)	0-7	133	11
	8-15	51	1

Table 4.2

### The Combination of 2x2 Tables

In one of the final sections of his 1954 paper, Cochran turns to the problem of developing a combined test of significance for the difference in proportions as displayed in a series of 2x2 tables.

Suppose that for the  $i$ th 2x2 table we have sample sizes  $n_{i1}$  and  $n_{i2}$ , observed difference in proportions  $d_i = x_{i1}/n_{i1} - x_{i2}/n_{i2}$ , and marginal (combined) proportion  $\hat{p} = (x_{i1} + x_{i2}) / (n_{i1} + n_{i2})$ . Then Cochran's test criterion is (in its 1 d.f.  $\chi^2$  version)

$$C^2 = \frac{(\sum w_i d_i)^2}{\sum w_i p(1-p)} \quad (5.1)$$

where

$$w_i = \frac{n_{i1} n_{i2}}{n_{i1} + n_{i2}} = \left( \frac{1}{\frac{1}{n_{i1}}} + \frac{1}{\frac{1}{n_{i2}}} \right)^{-1} \quad (5.2)$$

We now know that  $C^2$  is asymptotically equivalent to the UMP unbiased test for conditional independence (i.e. independence in each 2x2 table) versus the alternative of a constant odds-ratio (i.e. no-2nd-order interaction) across 2x2 tables (Birch, 1965).

To illustrate Cochran's test we apply it to the four 2x2 tables in Table 4.2:

$$C^2 = \frac{(10.5518)^2}{22.9886} = 4.85.$$

Since this is significant at the .05 level, we are able, with Cochran's 1 d.f. test, to detect the presence of the shift or regression alternative, that we found using computational more complex methods in the preceding section.

In 1959, Mantel and Haenszel independently proposed what is essentially Cochran's test, with two minor modifications (i) a "small sample" adjustment in the weights  $\{w_i\}$  in the denominator of  $C^2$ , and (ii) the use of a  $\frac{1}{2}$  continuity correction in the numerator. There has been a tendency for subsequent authors to refer to the Mantel-Haenszel test, and to ignore the fact that Cochran derived it first. It is a measure of Cochran's stature and temperament that he never took offense at this oversight. In fact, in the 7th edition of Snedecor and Cochran (1980), as in the preceding edition, he noted the importance of the two refinements in the Mantel-Haenszel version of his statistic.

#### 6. The Impact of Cochran's Research on Categorical Data Problems

It would be a mistake to assess the importance of Cochran's work in categorical data by simply looking at the new results he derived. In a sense, it is the impact of Cochran's work (a) on the research of others, and (b) on statistical practice, that he would have wanted us to examine.

As Colton (1981) has suggested, one measure of Cochran's impact on the research of others is the frequency with which some of his important contributions are still cited in the literature. Table 6.1 gives the number of citations listed in the Science Citation Index for 1965-1980 to those three of Cochran's on categorical data analysis which were the focus of our discussion in Sections 2-5. (Since the recommendations on small expectations in Cochran (1952, 1954) supercede those in his earlier papers, the earlier papers, though important, are no longer widely cited.) The remarkable feature discernable in Table 6.1 is that Cochran's articles, 25 to 30 years after their publication, are cited in the statistical literature more often than they were in the 1960's.

## Citation Counts

		1965-1969	1970-1974	1975-1979 <sup>1</sup>	1980
	Cochran (1950)	29	48	59	15
Paper	Cochran (1952)	19	18	36	5
	Cochran (1954)	67	159	196	34

Table 6.1  
Citation Counts from *Science Citation*  
Index for 1965-1980.

Snedecor and Cochran (1980) remains one of the most widely cited books in the entire scientific literature, and it serves as a methodological guide for statisticians and nonstatisticians alike. Throughout its discussion of categorical data analysis one can find, deftly woven between the basic methods and the informative examples and applications, many of Cochran's own recommendations and contributions. Thus we can expect Cochran's work to continue to influence the practice of scientists dealing with categorical data.

About a year prior to his death, in the winter of 1979, Bill Cochran visited the University of Minnesota to take part in a seminar on the contributions to statistics of Sir R.A. Fisher. In his lecture, Cochran (1980) made reference to his own 1940 paper on transformations for Poisson and binomial data, and the exchange he had with Fisher on the topic. That lecture was impressive for its modesty and the unpretentious manner with which Cochran put his own important contributions into perspective. His 1940 paper was clearly correct, and provided an innovative approach to a problem Fisher himself had posed. But when Fisher (1954) reviewed the various approaches to the analysis of non-normal data through the use of variance stabilizing transformations, he implied (incorrectly) that Cochran supported such an approach in preference to the maximum likelihood one. Cochran in fact had shown how to approximate the M.L. solution, but in his 1979 lecture, he remarked that

In retrospect I agree that Fisher's approach is superior. It specifies a definite mathematical model, and uses maximum likelihood estimation, recognized as preferable

<sup>1</sup> The 1978 figures are slightly understated, as they do not include December of that year.

to least squares estimation with non-normal data. My results were that analysis of variance on the transformed scale, with or without variance-stabilizing adjustments, agreed closely with the ML estimates, and was a good working method, particularly when extraneous variation is present so that the assumptions leading to Fisher's ML solutions do not apply.

When doing statistical research on a topic such as  $\chi^2$  tests for categorical data we tend to use the recent literature as a point of departure. Rereading Bill Cochran's papers on the analysis of categorical data, one realizes that we all have much more to gain by going back to several of his key papers on the topic, and reflecting on the wisdom contained therein, rather than relying on someone else's summary of their contents.

## REFERENCES

- Birch, M.W. (1965). "The detection of partial association. II: the general case." *J.R. Statist. Soc. B*, 27, 111-24.
- Bishop, Yvonne, M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass.: MIT Press.
- Cochran, W.G. (1936a). "Statistical analysis of field counts of diseased plants." *J.R. Statist. Soc. Suppl.* 3, 49-67.
- Cochran, W.G. (1936b). "The  $\chi^2$  distribution for the binomial and Poisson series with small expectations." *Ann. Eugen. Lond.* 7, 207-17.
- Cochran, W.G. (1940). "The analysis of variance when experimental errors follow the Poisson or binomial laws." *Ann. Math. Stat.* 11, 335-347.
- Cochran, W.G. (1942). "The chi-square correction for continuity." *Iowa State Coll. Jour. Sci.* 16, 421-436.
- Cochran, W.G. (1950). "The comparison of percentages in matched samples." *Biometrika* 37, 256-66.
- Cochran, W.G. (1952). "The  $\chi^2$  test of goodness of fit." *Ann. Math. Statist.* 23, 315-45.
- Cochran, W.G. (1954). "Some methods for strengthening the common  $\chi^2$  tests." *Biometrics* 10, 417-51.
- Cochran, W.G. (1955). "A test of a linear function of the deviations between observed and expected numbers." *J. Amer. Statist. Assoc.* 50, 377-97.
- Cochran, W.G. (1980). "Fisher and the analysis of variance." In S.E. Fienberg and D.V. Hinkley, R.A. Fisher: *An Appreciation*, Lecture Notes in Statistics, Vol. 1. New York: Springer-Verlag, 17-34.
- Colton, T. (1981). "Bill Cochran: His contributions to medicine and public health and some personal recollections." *Amer. Statist.* 35, 167-170.
- Conover, W.J. (1974). "Some reasons for not using the Yates continuity correction on 2 x 2 contingency tables (with Comments)." *J. Amer. Statist. Assoc.* 69, 374-82.
- Cox, D.R. (1970). *The Analysis of Binary Data*. London: Methuen.
- Duncan, O.D. (1982). "Rasch measurement in survey research: further examples and discussion." To appear in C.F. Turner and E. Martin (eds.) *Survey Measurement of Subjective Phenomena*, Vol. 2, Harvard Univ. Press.
- Fienberg, S.E. (1979). "The use of chi-squared statistics for categorical data problems." *J.R. Statist. Soc. B*, 41, 54-64.

Fisher, R.A. (1954). "The analysis of variance with various binomial transformations." *Biometrics* 10, 130-139.

Grizzle, J.E. (1967). "Continuity correction in the  $\chi^2$  test for 2 x 2 tables." *Amer. Statist.* 21, 28-32.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Lancaster, H.O. (1949). "The derivation and partition of  $\chi^2$  in certain discrete distributions." *Biometrika* 36, 117-29.

Larntz, K. (1978). "Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics." *J. Amer. Statist. Assoc.* 73, 253-63.

Mantel, N. & Haenszel, W. (1959). "Statistical aspects of the analysis of data from retrospective studies of disease." *J. Nat. Cancer. Inst.* 22, 719-48.

Plackett, R.L. (1964). "The continuity correction in 2 x 2 tables." *Biometrika* 51, 327-337.

Plackett, R.L. (1981). *The Analysis of Categorical Data*. 2nd edition. London: Griffin.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. (Reprinting of the 1960 book.) Univ. of Chicago Press.

Snedecor, G.W. and Cochran, W.G. (1980). *Statistical Methods*. 7th edition. Iowa State Univ. Press.

Stewart, Wm. (1981). Personal Communication.

Tjur, T. (1981). "A connection between Rasch's item analysis model and a multiplicative Poisson model." *Scan. J. Statistics* (in press).



SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report #222	2. GOVT ACCESSION NO. AD-A1113014	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) The Contributions of William Cochran to Categorical Data Analysis		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Stephen E. Fienberg		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE
		13. NUMBER OF PAGES 21
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The paper reviews the contributions of Wm. Cochran to the analysis of categorical data, focussing on: (a) the distribution of $\chi^2$ test statistics in the presence of small expectations, and the correction for continuity. (b) the Q-test for comparing percentages in matched samples. (c) methods for strengthening $\chi^2$ -tests. (d) the Cochran test for combining results from several 2x2 tables. Some links are made between Cochran's work and recent literature on the analysis of categorical data using loglinear models.		

DD FORM 1473

EDITION OF 1 NOV 65 IS OBSOLETE

GPO 770021-5601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

DATE  
FILMED

3-8